An acoustically-motivated spatial prior for under-determined reverberant source separation

Ngoc Q. K. Duong, Emmanuel Vincent, Rémi Gribonval

AN ACOUSTICALLY-MOTIVATED SPATIAL PRIOR FOR UNDER-DETERMINED REVERBERANT SOURCE SEPARATION

Ngoc Q. K. Duong, Emmanuel Vincent and Rémi Gribonval

INRIA, Centre de Rennes - Bretagne Atlantique
Campus de Beaulieu, 35042 Rennes Cedex, France
{qduong@irisa.fr, emmanuel.vincent@inria.fr, remi.gribonval@inria.fr}

ABSTRACT

We consider the task of under-determined reverberant audio source separation. We model the contribution of each source to all mixture channels in the time-frequency domain as a zero-mean Gaussian random vector with full-rank spatial covariance matrix. We introduce an inverse Wishart prior over the covariance matrices, whose mean is given by the theory of statistical room acoustics and whose variance is learned from training data. We then derive an Expectation-Maximization (EM) algorithm to estimate the model parameters in the Maximum A Posteriori (MAP) sense given prior knowledge about the microphone spacing and the source positions. This algorithm provides a principled solution to the well-known permutation problem and achieves better separation performance than other algorithms exploiting the same prior knowledge.

Index Terms— Under-determined convolutive source separation, full-rank spatial covariance, statistical room acoustics, inverse-Wishart prior.

1. INTRODUCTION

Under-determined audio source separation is the task of extracting J sources from a mixture signal consisting of $I < J$ channels. The $I \times 1$ mixture signal $\mathbf{x}(t)$ can be expressed as

$$\mathbf{x}(t) = \sum_{j=1}^{J} \mathbf{c}_j(t) \tag{1}$$

where $\mathbf{c}_j(t)$ is the spatial image of the j-th source, that is its contribution to all mixture channels [1]. For point sources in a reverberant setting, this quantity is equal to the convolution of the original source signals $s_j(t)$ by mixing filters modeling acoustic propagation from the source to the microphones.

Most state-of-the-art approaches operate in the time-frequency domain by means of the Short-Time Fourier Transform (STFT). Under a narrowband assumption, time-domain convolution is approximated by complex-valued multiplication in the frequency domain such that the STFT coefficients of each source image are given by $\mathbf{c}_j(n,f) = \mathbf{h}(f)s_j(n,f)$ in time frame n and frequency bin f where $\mathbf{h}_j(f)$ is the Fourier transform of the mixing filters. The sources are then estimated under additional sparsity assumptions [2, 3, 1].

Recently, a distinct local Gaussian framework has emerged [4, 5] whereby $\mathbf{c}_j(n,f)$ are modeled as zero-mean Gaussian random variables with covariance matrix

$$\mathbf{\Sigma}_j(n,f) = v_j(n,f)\,\mathbf{R}_j(f) \tag{2}$$

where $v_j(n,f)$ are scalar time-varying *variances* encoding the spectro-temporal power of the source and $\mathbf{R}_j(f)$ are $I \times I$ full-rank time-invariant *spatial covariance matrices* encoding its spatial position and spatial spread. This framework was shown to better model the convolutive mixing process and to improve separation performance compared to state-of-the-art approaches based on the narrowband approximation [4]. However, the Maximum Likelihood (ML) parameter estimation algorithm proposed in [4] remains sensitive to initialization and relies on the post-processing algorithm in [3] to solve the permutation problem, that is to align the order of the sources across frequency.

In this paper, we introduce an inverse Wishart prior over the spatial covariance matrices $\mathbf{R}_j(f)$, whose mean is given by the theory of statistical room acoustics and whose variance is learned from training data, and show that it is especially accurate for large reverberation times. We then derive an Expectation-Maximization (EM) algorithm to estimate the model parameters in the Maximum A Posteriori (MAP) sense given prior knowledge about the microphone spacing and the source positions. This algorithm offers an acoustically principled solution to the estimation of the model parameters and to the permutation problem and may be used in situations with known geometric setting, for instance in a formal meeting or in a car environment. Most importantly, it provides a proof of concept of the benefit of the proposed prior towards its future use in a blind source separation context.

The structure of the rest of the paper is as follows. We introduce the proposed spatial prior based on statistical room acoustics in Section 2 and address MAP estimation of the model parameters in Section 3. We then learn and discuss the prior variance hyper-parameter and provide experimental results to confirm the effectiveness of the proposed approach in Section 4. We conclude in Section 5.

2. ACOUSTICALLY-MOTIVATED SPATIAL PRIOR

Under the mixing model (1) and the parameterization (2), assuming that the sources are uncorrelated, the vector of STFT coefficients of the mixture signal $\mathbf{x}(n,f)$ is zero-mean Gaussian with covariance matrix

$$\mathbf{\Sigma_x}(n,f) = \sum_{j=1}^{J} v_j(n,f)\mathbf{R}_j(f). \tag{3}$$

The log-likelihood is then given by

$$\log \mathcal{L} = -\sum_{n,f} \log\det(\pi\,\mathbf{\Sigma_x}(n,f)) + \mathrm{tr}(\mathbf{\Sigma_x}^{-1}(n,f)\widehat{\mathbf{\Sigma}}_{\mathbf{x}}(n,f)) \tag{4}$$

where $\det(.)$ denotes the determinant of a square matrix and $\widehat{\mathbf{\Sigma}}_{\mathbf{x}}(n,f)$ the empirical mixture covariance matrix as defined in [5]. We now focus on designing a suitable prior distribution over $\mathbf{R}_j(f)$.

In [3], a quadratic cost function was proposed under the narrowband assumption to minimize the difference between phase- and amplitude-normalized versions of the mixing vectors $\mathbf{h}_j(f)$ in a reverberant environment and their values in an anechoic environment. This function amounts to a Gaussian prior over the normalized mixing vectors whose mean is given by the anechoic model. Although the benefit of this prior was demonstrated for source separation, the accuracy of the chosen mean and the resulting variance were not investigated.

In the following, we model $\mathbf{R}_j(f)$ as

$$p(\mathbf{R}_j(f)) = \mathcal{IW}\big(\mathbf{R}_j(f)|\mathbf{\Psi}_j(f), m\big) \tag{5}$$

where

$$\mathcal{IW}(\mathbf{R}|\mathbf{\Psi}, m) = \frac{|\mathbf{\Psi}|^m|\mathbf{R}|^{-(m+I)}e^{-\mathrm{tr}(\mathbf{\Psi}\mathbf{R}^{-1})}}{\pi^{I(I-1)/2}\prod_{i=1}^{I}\Gamma(m-i+1)} \tag{6}$$

is the inverse Wishart density over a Hermitian positive definite matrix $\mathbf{R}$ with positive definite inverse scale matrix $\mathbf{\Psi}$, m degrees of freedom and mean $\mathbf{\Psi}/(m-I)$ [6], with $\mathrm{tr}(.)$ denoting the trace of a square matrix and Γ the gamma function. This distribution, its mean, and its variance exists for $m > I-1$, $m > I$, and $m > I+1$ respectively. We chose this prior as the conjugate prior for the likelihood of the considered Gaussian observation model, so that it results in closed-form parameter update equations.

According to the theory of statistical room acoustics, for a given microphone spacing and source position relative to the microphones, the mean spatial covariance matrix over all possible microphone positions is equal to [7]

$$\frac{\mathbf{\Psi}_j(f)}{m-I} = \mathbf{a}_j(f)\mathbf{a}_j^H(f) + \sigma^2_{\mathrm{rev}}\mathbf{\Omega}(f) \tag{7}$$

where the mixing vector $\mathbf{a}_j(f)$ models the direct path from the source to the microphones, σ^2_{rev} denotes the mean power of echoes and reverberation and $\mathbf{\Omega}(f)$ is the normalized covariance matrix of a diffuse noise. The entries of $\mathbf{a}_j(f)$ and $\mathbf{\Omega}(f)$ can be computed from the geometric setting as

$$a_{ij}(f) = \frac{1}{\sqrt{4\pi}r_{ij}}e^{-2i\pi f\frac{r_{ij}}{c}} \tag{8}$$

$$\mathbf{\Omega}_{ii'}(f) = \mathrm{sinc}(2\pi f d_{ii'}/c) \tag{9}$$

where r_{ij} is the distance from the j-th source to the i-th microphone, $d_{ii'}$ the microphone spacing, c the sound velocity and $\mathrm{sinc}(.) = \sin(.)/(.)$. Considering the special case of a parallelepipedic room with dimensions L_x, L_y, L_z,

$$\sigma^2_{\mathrm{rev}} = \frac{4\beta^2}{\mathcal{A}(1-\beta^2)} \tag{10}$$

where $\mathcal{A}$ is the total wall area and β the wall reflection coefficient computed from the room reverberation time T_{60} via Eyring's formula [7]

$$\beta = \exp\left\{-\frac{13.82}{(\frac{1}{L_x}+\frac{1}{L_y}+\frac{1}{L_z})cT_{60}}\right\}. \tag{11}$$

The proposed model (5) extends the so-called full-rank direct+diffuse model formerly introduced in [4] by allowing deviations of the spatial covariance matrices around their mean (7) controlled by the number of degrees of freedom m. Such deviations occur for instance when the source or the microphones are close to the walls, resulting in a strong directional early echo. The value of m is learned from training data and discussed in Section 4.1.

3. MAP ESTIMATION OF MODEL PARAMETERS

Given the prior hyper-parameters $\mathbf{\Psi}_j(f)$ and m, we now estimate the model parameters $\theta = \{v_j(n,f), \mathbf{R}_j(f),\ \forall j,n,f\}$ in the MAP sense. We consider an EM algorithm, which is a well-known approach for ML or MAP parameter estimation in statistical Gaussian models. The *complete data* is chosen as $\{\mathbf{c}_j(n,f)\ \forall j,n,f\}$, that is the set of STFT coefficients of all source images in all time-frequency bins.

In the E-step of the algorithm, the expected covariance matrices $\widehat{\mathbf{\Sigma}}_j(n,f)$ are updated similarly as in [5] using the Wiener filters $\mathbf{W}_j(n,f)$

$$\mathbf{W}_j(n,f) = \mathbf{\Sigma}_j(n,f)\mathbf{\Sigma}_{\mathbf{x}}^{-1}(n,f) \tag{12}$$

$$\widehat{\mathbf{\Sigma}}_j(n,f) = \mathbf{W}_j(n,f)\widehat{\mathbf{\Sigma}}_{\mathbf{x}}(n,f)\mathbf{W}_j^H(n,f) + (\mathbf{I}-\mathbf{W}_j(n,f))\mathbf{\Sigma}_j(n,f) \tag{13}$$

where $\mathbf{I}$ is the $I\times I$ identity matrix, $\mathbf{\Sigma}_j(n,f)$ is defined in (2) and $\mathbf{\Sigma}_{\mathbf{x}}(n,f)$ in (3).

In the M-step of the algorithm, the auxiliary function Q defined in the MAP sense as

$$Q_{MAP}(\theta|\theta^{\mathrm{old}}) = \sum_{j,f}\Big(\sum_n \log p\big(\mathbf{c}_j(n,f)|\mathbf{0},\mathbf{\Sigma}_j(n,f)\big) + \gamma\log p\big(\mathbf{R}_j(f)|\mathbf{\Psi}_j(f),m\big)\Big) \tag{14}$$

is maximized with respect to the parameters, where γ is a tradeoff hyper-parameter determining the contribution of the prior, $p\big(\mathbf{R}_j(f)|\mathbf{\Psi}_j(f), m\big)$ is defined in (6), $\mathbf{\Sigma}_j(n,f)$ in (2), and $\log p(\mathbf{c}_j(n,f)|\mathbf{0}, \mathbf{\Sigma}_j(n,f)) = -\log\det(\pi\,\mathbf{\Sigma}_j(n,f)) - \mathrm{tr}(\mathbf{\Sigma}_j^{-1}(n,f)\widehat{\mathbf{\Sigma}}_j(n,f))$. By computing the partial derivative of $Q_{MAP}(\theta|\theta^{\text{old}})$ with respect to $v_j(n,f)$ and to each entry of $\mathbf{R}_j(n,f)$ and equating it to zero, we obtain the update rules

$$v_j(n,f) = \frac{1}{I}\,\mathrm{tr}\left(\mathbf{R}_j^{-1}(f)\widehat{\mathbf{\Sigma}}_j(n,f)\right) \tag{15}$$

$$\mathbf{R}_j(f) = \frac{1}{\gamma(m+I)+N}\Big(\gamma\mathbf{\Psi}_j(f) + \sum_{n=1}^{N}\frac{\widehat{\mathbf{\Sigma}}_j(n,f)}{v_j(n,f)}\Big) \tag{16}$$

with N denoting the number of time frames.

4. EXPERIMENTAL RESULTS

4.1. Computation and analysis of the prior variance

In order to learn the number of degrees of freedom of the proposed prior (5), we generated room impulse responses via the image method for 20 random source positions for each of 20 random microphone pair positions using the Roomsim toolbox[1]. The room dimensions were $4.45 \times 3.55 \times 2.5$ m and the microphone spacing and the distance from sources to center of the microphone pair were fixed to 5 cm and 50 cm, respectively. Four different reverberation times were considered: $T_{60} = 50$, 130, 250 and 500 ms. Source images were computed by convolving a 10 s male speech source with the simulated impulse responses. This resulted in a total of 400 source image signals indexed by p for each reverberation time. For each of these source images, the spatial covariance matrix $\mathbf{R}_p(f)$ was computed at each frequency f in the ML sense by alternatingly applying (15) and (16) with $\gamma = 0$.

Since $\mathbf{R}_p(f)$ can be measured only up to an arbitrary scaling factor α, assuming that α is distributed according to a Jeffreys prior, the number of degrees of freedom m may be estimated in the ML sense by maximizing

$$\log\mathcal{L} = \sum_p\sum_f \log\int_0^\infty p\big(\mathbf{R}_p(f)|\alpha, \mathbf{\Psi}_p(f), m\big)p(\alpha)d\alpha \tag{17}$$

where $p\big(\mathbf{R}_p(f)|\alpha, \mathbf{\Psi}_p(f), m\big) = J_\alpha\mathcal{IW}\big(\alpha\mathbf{R}_p(f)|\mathbf{\Psi}_p(f), m\big)$, $J_\alpha = \alpha^{I^2}$ is the Jacobian of the scaling transform, $p(\alpha) = 1/\alpha$, and $\mathbf{\Psi}_p(f)$ was computed by (7) for each geometry setting p. By first computing the integral and then using Matlab's fmincon Newton-based optimizer, the optimal value of m was found. This value is shown in Table 1 together with the mean power σ^2_{rev} of echoes and reverberation computed by (10), which are both determined by the reverberation time.

[1] http://www.irisa.fr/metiss/members/evincent/Roomsimove.zip

T_{60}	50 ms	130 ms	250 ms	500 ms
m	2.1	2.8	4.2	6.4
σ^2_{rev}	0.011	0.057	0.131	0.287

Table 1. Learned value of m and predicted value of σ^2_{rev}.

As expected, σ^2_{rev} strongly increases with reverberation, such that the direct-to-reverberant energy ratio is 14 dB lower when $T_{60} = 500$ ms than when $T_{60} = 50$ ms. More surprisingly, the variance of the prior, which is inversely related to m [6], decreases with reverberation time while the empirical variance (not shown in the Table) follows the opposite trend. This observation suggests that the inverse Wishart prior is inappropriate for small reverberation where the early echoes and later echoes do not form a diffuse soundfield.

4.2. Source separation performance

In order to evaluate the separation performance, we generated the impulse responses from $J = 4$ sources to $I = 2$ microphones for the same room, the same reverberation times and the same microphone spacing and distance from the sources to the center of the microphone pair as above. The source directions-of-arrival are 20^o, 80^o, 120^o and 150^o. Three mixtures were generated by convolving 10 s speech sources (male voice, female voice, and mixed male and female voices) sampled at 16 kHz with the simulated impulse responses. We compare the proposed MAP-based algorithm to the ML-based algorithms in [4], when $\mathbf{R}_j(f)$ was initialized either blindly as in [4] (named Blind init. full-rank ML) or the from geometric setting by $\mathbf{\Psi}_j(f)$ (named Geom. init full-rank ML). We also compute the baseline separation offered by binary masking [2] where the mixing vectors were fixed to the first eigenvector of $\mathbf{\Psi}_j(f)$ (named Geom. init. binary masking). Note that we do not consider rank-1 model-based and ℓ_p-norm minimization based algorithms here since they were shown in [4] to be outperformed by binary masking in moderate to high reverberation condition. The STFT was computed with a sine window of length 1024 and the number of EM iterations was 20. The trade-off parameter γ does not significantly affect the result but we observed that $\gamma = 50$ is globally a good choice. The separation performance is evaluated in terms of signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR), and signal-to-artifact ratio (SAR) criteria expressed in decibels (dB), as defined in [8], averaged for all sources and all mixtures, and shown in Fig. 1, Fig. 2, and Fig. 3, respectively.

It is not surprised that binary masking results in the best SIR among other algorithms but as contrary its SAR is very poor. The proposed MAP-based algorithm provides higher SIR and moderate SAR compared to the full-rank ML based approaches. Overall, the MAP-based algorithm outperforms all other algorithms for all considered reverberation times in term of SDR criterion, which measures the overall distor-

tion, confirming the benefit of the proposed approach. For instance, at $T_{60} = 130$ ms the proposed MAP based algorithm offers 1.2 dB, 2.1 dB and 1.2 dB higher SDR than that achieved by Geom. init full-rank ML, Bind. init full-rank ML and Geom. init binary masking, respectively.

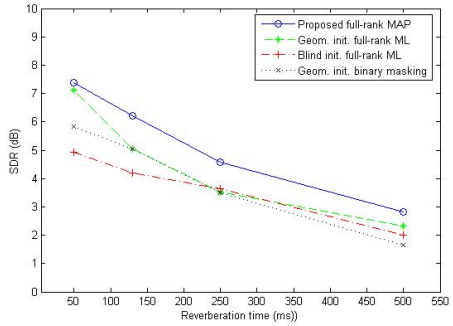

Fig. 1. Averaged SDR as function of reverberation time.

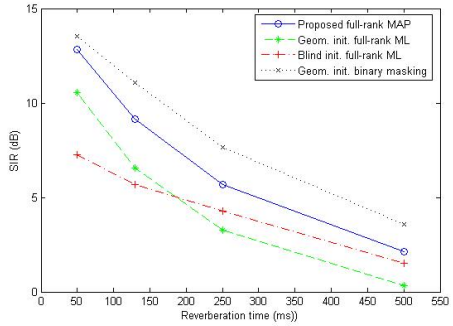

Fig. 2. Averaged SIR as function of reverberation time.

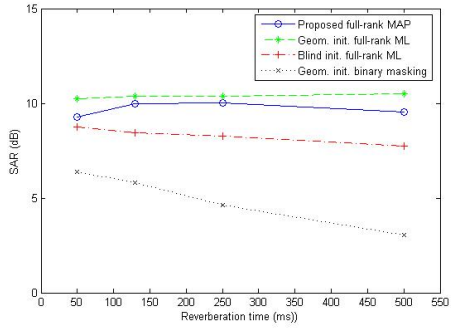

Fig. 3. Averaged SAR as function of reverberation time.

5. CONCLUSION

In this paper, we addressed the reverberant source separation problem using full-rank spatial covariance model where an acoustically-motivated spatial prior from the theory of statistical room acoustics was introduced. Given prior knowledge about the geometric setting, we derived the estimation of the model parameters in the MAP sense. Experimental results over several reverberation conditions confirm the benefit of the proposed approach compared to other algorithms exploiting the same prior knowledge.

6. ACKNOWLEDGMENT

This work is part of the i3DMusic project funded by Oseo.

7. REFERENCES

[1] E. Vincent, M. G. Jafari, S. A. Abdallah, M. D. Plumbley, and M. E. Davies, "Probabilistic modeling paradigms for audio source separation," in *Machine Audition: Principles, Algorithms and Systems*, pp. 162–185. IGI Global, 2010.

[2] Ö. Yılmaz and S. T. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[3] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1592–1604, 2007.

[4] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 18, no. 7, pp. 1830–1840, Sep. 2010.

[5] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using local observed covariance and auditory-motivated time-frequency representation," in *Proc. LVA/ICA*, Sep. 2010, pp. 73–80.

[6] D. Maiwald and D. Kraus, "Calculation of moments of complex Wishart and complex inverse-Wishart distributed matrices," *IEE Proceedings on Radar, Sonar and Navigation*, vol. 147, pp. 162–168, 2000.

[7] Tony Gustafsson, B. D. Rao, and M. Trivedi, "Source localization in reverberant environments: Modeling and statistical analysis," *IEEE Trans. on Speech and Audio Processing*, vol. 11, pp. 791–803, Nov 2003.

[8] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J.P. Rosca, "First Stereo Audio Source Separation Evaluation Campaign: Data, algorithms and results," in *Proc. ICA*, 2007, pp. 552–559.